

Application of iterated Bernstein operators to distribution function and density approximation

Claude Manté

Aix-Marseille Université, MIO, UMR C.N.R.S. 7294, Modelling, Statistics and Biological
Systems Data Analysis Team, Campus de Luminy, Case 901, F13288 Marseille Cedex 09,
France

☎ (+33) 491 829 127 fax: (+33) 491 829 119
email: claudio.mante@univ-amu.fr

Abstract

We propose a density approximation method based on Bernstein polynomials, consisting in superseding the classical Bernstein operator by a convenient number I^* of iterates of a closely related operator. We mainly tackle two difficulties met in processing real data, sampled on some mesh X_N . The first one consists in determining an optimal sub-mesh X_{K^*} , in order that the operator associated with X_{K^*} can be considered as an authentic Bernstein operator (necessarily associated with a uniform mesh). The second one consists in optimizing I in order that the approximated density is *bona fide* (positive and integrates to one). The proposed method is tested on two benchmarks in Density Estimation, and on a grain-size curve.

Keywords: Non-parametric density estimator, Bernstein polynomials, *Bona fide* density, Optimal mesh, Hausdorff metric.

1. Introduction

S. Bernstein introduced the polynomials that bear his name in his famous constructive proof of the Stone-Weierstrass theorem [5], using probabilistic arguments to prove the uniform convergence of his approximation. This dual belonging to the fields of Approximation Theory and Statistics benefits to both communities : statistical arguments are often used in Approximation Theory [20, 36], while the attractive approximation properties of Bernstein polynomials prompted statisticians to introduce them in Density Estimation [2, 8, 14, 28].

In this paper, we propose a density approximation method also based on both disciplines approaches. While classical Density Estimation deals with empirical distribution functions (e.d.f.s), this method is designed for processing discretized distribution functions (d.d.f.s), defined hereunder. By convention, all these distribution functions will be defined on the unit interval, except in the illustrating applications.

Definition 1. An e.d.f. F_K is built from the sorted sample of values of some random variable X , $\{0 \leq x_{(1)} < x_{(2)} \dots, < x_{(K-1)} \leq 1\}$; F_K is the step function associated with the jump set

$$S_K := \left\{ (x_{(0)} = 0, 0), \left(x_{(1)}, \frac{1}{K}\right), \dots, \left(x_{(K-1)}, \frac{K-1}{K}\right), (x_{(K)} = 1, 1) \right\}.$$

In contrast with e.d.f.s, whose jump sets have random abscissas and deterministic ordinates, d.d.f.s have deterministic abscissas and random ordinates.

Definition 2. A d.d.f. is built from a finite list $F_K := \{(x_i, F(x_i)) : 0 \leq i \leq K\}$ of values of some increasing function F sampled on a fixed mesh $x_0 = 0 \leq x_1 < \dots, < x_{K-1} \leq 1 = x_K$, whose image is in the unit interval. The associated step function will be denoted F_K too.

An e.d.f. possesses of course the same structure as a d.d.f., but there is a lot of data used in geoscience, medicine or other fields (grain-size curves, pore-size curves, thermal remanent magnetization curves, biomass spectra, growth curves, ...), which are d.d.f.s, not e.d.f.s. They belong to the same class of random functions as e.d.f.s, but the individual characteristic (size, ...) of each object (particle, pore, ...) is unknown; furthermore, in some cases, these objects are not even counted (ponderal grain-size curves, for instance). That is why we will speak of density approximation, not density estimation. But even though d.d.f.s do not consist in e.d.f.s, they can be associated too with probability measures. In a series of works [24, 25, 26, 27], we proposed a convenient functional method designed for the exploratory analysis of a S -sample $\{\nu_1, \dots, \nu_S\}$ of measures. Roughly speaking, it consists in performing the Principal Component Analysis (PCA) of the Radon-Nikodym densities $\{d\nu_1/d\mu, \dots, d\nu_S/d\mu\}$ where μ is a reference probability dominating all these measures. The point is that we have shown that, ideally, the S d.d.f.s should be discretized on a mesh consisting in fractiles of the reference distribution function F_μ (μ -optimal meshes). Thus, a practical issue for PCA of such data (grain-size curves, for instance) is the “fractiles problem”: given a reference distribution F_μ sampled on some imposed mesh X_N (giving rise to the observed d.d.f. F_N) how shall we extract a sub-mesh X_K consisting (at last approximately) in fractiles of F_μ ?

This paper is organized as follows. In Section 2, we recall results from the Approximation Theory literature, essential for our purpose, and prove that the rate of point-wise convergence of the standard Bernstein approximation for a distribution function sampled on K positions is typically $\frac{1}{\sqrt{K}}$. The Sevy’s method [31, 30] for accelerating the convergence of Bernstein approximations, based on iterations of the associated operator, is reminded there. In Section 3, we review recent works about the use of Bernstein polynomials in Density Estimation, and about the numerical problems they pose. A topic of cardinal importance [29] consists in optimizing the number of iterations in the Sevy’s scheme. This point is tackled in the last part of Section 4. The first part of Section 4 is dedicated to the problem of optimal sub-sampling: in order to compute the Bernstein approximation of degree K of some continuous function F , we need a

uniform sample of F , of size $K + 1$. But in general, either the sampling mesh is not uniform, or the number of sampled values is excessive : for instance, if $K + 1 = 100$, the degree of the polynomial is 99! We propose tools for extracting from the original mesh sub-meshes well-suited for our purpose. The method is then tested on two benchmarks in Density Estimation and on a grain-size curve ; an heuristics for the fractiles problem is proposed in this section, based on our density approximation. At last, Section 5 is devoted to discussion.

2. The Approximation Theory viewpoint

First of all, let us briefly recall the proof of the Stone-Weierstrass theorem by Bernstein [5, 22]. Let $F \in C^0[0, 1]$, and A be an event whose success probability is $x \in [0, 1]$. Bernstein supposes a gambler performs K independent Bernoulli trials, and wins $F(\frac{J}{K})$ if A has been drawn J (random) times. His average gain is then :

$$B_K[F](x) := \sum_{j=0}^K F\left(\frac{j}{K}\right) W_{K,j}(x) \quad (1)$$

with $W_{K,j}(x) := \binom{K}{j} x^j (1-x)^{K-j}$. Then, since $\frac{J}{K} \xrightarrow{\mathcal{P}} x$ (Bernoulli's law of large numbers), $\|B_K[F] - F\| \xrightarrow{K \rightarrow \infty} 0$, where $\|G\| := \sup_{x \in [0,1]} |G(x)|$.

Of course, if F is positive, $B_K[F]$ is so, but a number of researchers established further attractive properties of B_K . For instance, if $F(x) \in [a, b]$, $B_K[F](x) \in [a, b]$; if $F \in C^p[0, 1]$, $\|B_K[F]^{(q)} - F^{(q)}\| \xrightarrow{K \rightarrow \infty} 0 \quad \forall q \leq p$; if F is monotone, $B_K[F]$ is monotone and if F is convex, $B_K[F]$ is so (see *e.g.* [12]). But “there is a price that must be paid for these beautiful approximation properties : the convergence of the Bernstein polynomials is very slow” [12] . Nevertheless, the Voronovsky theorem [12] proves that the rate of point-wise convergence is $1/K$ at points where F is sufficiently regular ; if F is of class $C^m[0, 1]$ ($m \geq 2$), this convergence is uniform at the rate $1/K$. When F is only continuous, Zeng and Piriou [36] proved that the rate of uniform convergence of $B_K[F]$ towards F is bounded by $1.25 \omega(K^{-1/2}, F)$. They also gave [36] an upper bound of the point-wise rate of convergence of $B_K[F]$ for bounded variation functions on $[0, 1]$:

Theorem 3. (Zeng and Piriou [36]) *Let F be of bounded variation on $[0, 1]$. Then, for every x in $[0, 1]$ and $K \geq 3$ we have:*

$$\left| B_K[F](x) - \frac{F(x^+) + F(x^-)}{2} \right| \leq \frac{3}{Kx(1-x) + 1} \mathcal{V}(F, x, K) + \frac{2}{\sqrt{Kx(1-x) + 1}} \mathcal{C}(F, x, K)$$

where $\mathcal{C}(F, x, K) = 0$ if F is continuous at x , and $\mathcal{V}(F, x, K) = \sum_{k=1}^K V_{x-x/\sqrt{k}}^{x+(1-x)/\sqrt{k}}(g_x)$,

with $g_x(t) = \begin{cases} F(t) - F(x^+) & \text{if } t \in]x, 1] \\ 0 & \text{if } t = x \\ F(t) - F(x^-) & \text{if } t \in [0, x[\end{cases}$ and $V_a^b(g_x)$ denotes the total variation of g_x on $[a, b]$.

In the particular case of d.f.s, we will give hereunder a better upper bound ; but we need first a preliminary result.

Lemma 4. *If n is large enough,*

$$\sum_{m=2}^{n-1} \sqrt{m} \left(1 - \sqrt{\frac{m}{m+1}} \right) = \sqrt{n} - \frac{4\sqrt{2}}{3} + \frac{3\sqrt{\frac{1}{n}}}{4} + O\left[\frac{1}{n-2}\right].$$

Proof. Consider first the functions $w(x) := \sqrt{x} \left(1 - \sqrt{\frac{x}{x+1}} \right)$, $w_n(x) := (n-2) \sqrt{(n-2)x+1} \left(1 - \sqrt{\frac{x(n-2)+1}{(n-2)x+2}} \right)$, and the integral

$$\begin{aligned} \int_1^{n-1} w(x) dx &= \int_0^1 w_n(x) dx = -\frac{2}{3} \left(1 + \sqrt{2} + \sqrt{n-1} - 3\sqrt{n} - n\sqrt{n-1} + n^{3/2} \right) \\ &= \sqrt{n} - \frac{2}{3} \left(1 + \sqrt{2} \right) + \frac{\sqrt{\frac{1}{n}}}{4} + O\left[\frac{1}{n}\right]^{3/2}. \end{aligned}$$

Since w_n is of bounded variation, the Theorem 1(b) of Chui [9] leads to : $\sum_{m=2}^{n-1} \sqrt{m} \left(1 - \sqrt{\frac{m}{m+1}} \right) - \int_0^1 w_n(x) dx = O\left[\frac{1}{n-2}\right]$. The lemma is then proved by keeping only the dominant terms of the integral \square

Remark. One can easily verify that the above approximation is indeed very good for any $n \geq 3$; moreover, notice that since $\sqrt{x} \left(1 - \sqrt{\frac{x}{x+1}} \right)$ is decreasing, the sum is slightly lower than the integral.

We can now prove a corollary of Theorem 3, designed for distribution functions.

Corollary 5. *Let F be an absolutely continuous d.f. on $[0, 1]$ associated with some probability P with $f := \frac{dP}{d\lambda}$, and suppose $\|f\| < +\infty$. Then, for $K \geq 3$*

$$|B_K[F](x) - F(x)| \leq \frac{6\|f\|}{x(1-x) + 1/K} \left(\frac{1}{\sqrt{K}} - O\left[\frac{1}{K-2}\right] \right).$$

Proof. Since F is continuous, $g_x(t) = F(t) - F(x) \forall x \in [0, 1]$ and, because F is a d.f., we can write :

$$\begin{aligned} V_{x-x/\sqrt{k}}^{x+(1-x)/\sqrt{k}}(g_x) &= F\left(x + (1-x)/\sqrt{k}\right) - F\left(x - x/\sqrt{k}\right) \\ &= \int_{I_k(x)} f(t) dt = P(I_k(x)) \end{aligned}$$

with $I_k(x) := [x - x/\sqrt{k}, x + (1-x)/\sqrt{k}]$. Notice that

$$I_K(x) \subsetneq I_{K-1}(x) \subsetneq \cdots \subsetneq I_1(x) = [0, 1].$$

As a consequence $\mathcal{V}(F, x, K) = K P(I_K(x)) + \sum_{k=1}^{K-1} (K-k) P(C_{K-k}(x))$, where

$C_{K-k}(x) = I_{K-k}(x) \cap \overline{I_{K-k+1}(x)}$ is a “crown” around x , with two disjoint components : a “left” one, $C_{K-k}^L(x) = [x - x/\sqrt{K-k}, x - x/\sqrt{K+1-k}]$, and a “right” one : $C_{K-k}^R(x) = [x + (1-x)/\sqrt{K+1-k}, x + (1-x)/\sqrt{K-k}]$. Since F is continuous, $P(I_K(x)) = \frac{F(x_K)}{\sqrt{K}}$ for some $x_K \in I_K(x)$; we have also : $\int_{C_{K-k}^L(x)} f(t)dt = P(C_{K-k}^L(x)) = x \Delta_K^k f(x_{K-k}^L)$, for some $x_{K-k}^L \in C_{K-k}^L(x)$ and $\Delta_K^k := \left(\frac{1}{\sqrt{K-k}} - \frac{1}{\sqrt{K-k+1}} \right) \leq 1 - 1/\sqrt{2}$. Similarly, $\int_{C_{K-k}^R(x)} f(t)dt = P(C_{K-k}^R(x)) = (1-x) \Delta_K^k f(x_{K-k}^R)$, for some $x_{K-k}^R \in C_{K-k}^R(x)$.

We obtain at last :

$$\begin{aligned} \mathcal{V}(F, x, K) &= \sqrt{K} f(x_K) + \sum_{k=1}^{K-2} (K-k) \Delta_K^k \frac{x f(x_{K-k}^L) + (1-x) f(x_{K-k}^R)}{2} + \\ &\quad \left(1 - 1/\sqrt{2}\right) \frac{x f(x_1^L) + (1-x) f(x_1^R)}{2} \\ &= \sqrt{K} f(x_K) + \sum_{j=2}^{K-1} \sqrt{j} \left(1 - \sqrt{\frac{j}{j+1}}\right) \frac{x f(x_j^L) + (1-x) f(x_j^R)}{2} + \\ &\quad \left(1 - 1/\sqrt{2}\right) \frac{x f(x_1^L) + (1-x) f(x_1^R)}{2} \end{aligned}$$

and finally, since f is bounded and because of lemma 4 :

$$\frac{3}{Kx(1-x)+1} \mathcal{V}(F, x, K) \leq \frac{6 \|f\|}{x(1-x)+1/K} \left(\frac{1}{\sqrt{K}} - O\left[\frac{1}{K}\right] \right)$$

□

Thus, the rate of point-wise convergence of Bernstein approximations of d.f.s is typically $O\left[\frac{1}{\sqrt{K}}\right]$ on $]0, 1[$. Remember now that, in order to accelerate the convergence of Bernstein approximations, Sevy [31, 30] proposed to substitute the iterated operator $\mathfrak{I}_K^I := \left(1 - (1 - B_K)^I\right)$ to B_K . This method has been recently re-discovered by Sahai [29], who noticed that one can write $C^0[0, 1] \ni F = B_K[F] + E$, where $E \in C^0[0, 1]$ is an unknown error which can be approached by $B_K[E]$. Then, $B_K[F] + B_K[E] = \left(1 - (1 - B_K)^2\right)[F]$ is a better approximation of F than $B_K[F]$, and so on... Sevy [30] proved that if $F \in C^{m+k}[0, 1]$, and if the number of iterations is $I > m$, $\left\| (\mathfrak{I}_K^I[F])^{(k)} - F^{(k)} \right\| =$

$o(K^{-\frac{\alpha}{2}})$; thus, if $F \in C^1[0, 1]$, for $I \geq 1$, $\mathfrak{J}_K^I[F]$ uniformly converges towards F at the rate $o\left[\frac{1}{\sqrt{K}}\right]$; if $F \in C^2[0, 1]$, for $I \geq 2$, $\mathfrak{J}_K^I[F]$ uniformly converges towards F at the rate $o\left[\frac{1}{K}\right]$, etc... Sevy also proved the following result :

Theorem 6. (*Sevy [31], see also Cooper and Waldron [10]*) For fixed $K \geq 1$ and any function F defined on $[0, 1]$,

$$\|\mathfrak{J}_K^I[F] - L_K[F]\| \xrightarrow{I \rightarrow \infty} 0 \quad (2)$$

where $L_K[F]$ is the Lagrange polynomial interpolating F at the $K+1$ equidistant nodes $\{\frac{k}{K}, 0 \leq k \leq K\}$.

This theorem will enable us in Section 4.3 to find a satisfactory trade-off between $B_K[F]$ (generally too smooth) and the Lagrange interpolation polynomial, which is notoriously a bad approximate [13, 22].

3. Bernstein polynomials for Density Estimation

Bernstein polynomials were introduced in Density Estimation according to two different strategies : the direct one consists in deriving an estimator of the d.f., while the indirect one goes through the estimated quantile function. In this section, N denotes the size of the studied sample.

3.1. The direct estimator

Let $F_N(x)$ be the e.d.f. associated with a N -sample of some real random variable X supported by $[0, 1]$, and $m < N$. Babu et al. [2] proposed an estimator $\tilde{F}_{N,m}$ of the distribution function F of X , consisting in smoothing F_N :

$$\tilde{F}_{N,m}(x) := \sum_{k=0}^m F_N\left(\frac{k}{m}\right) W_{m,k}(x) = B_m[F_N]. \quad (3)$$

They proved its almost sure convergence when the associated d.f. is continuous; moreover, they gave conditions under which its rate of stochastic convergence can be determined, as well as the rate of convergence of the associated density estimator $\tilde{f}_{N,m}(x) := \frac{d}{dx} \tilde{F}_{N,m}(x)$. The properties of $\tilde{f}_{N,m}$ were also investigated by Bouezmarni and Rolin [8], who proved not only its convergence under weaker conditions, but also that the estimation of a density supported by $[0, 1]$ and unbounded at $x = 0$ is itself asymptotically unbounded at $x = 0$ (another type of shape-preserving property).

3.2. The indirect estimator

In an interesting paper, de Bruin et al. [14] proposed another estimator of f . It consists in the inverse of the derivative of the increasing function :

$$H_N(p) := \sum_{i=0}^{N+1} x_{(i)} W_{N+1,i}(p).$$

Indeed, since $x_{(i)}$ is an estimate of $F^{-1}(i/(N+1))$, according to (1), $H_N(p) = B_{N+1}[F^{-1}](x)$ estimates the quantile function of X . Consequently, they estimate $f(x)$ by $\frac{d}{dx}H_N^{-1}(x)$.

It is noteworthy that their paper evades several practical problems. First, if N is large (100, say) $H_N(p)$ will be a polynomial of degree $N+1$ - not easily manageable with a computer! Moreover, in such cases, finding the root of the polynomial equation $x = H_N(p_x)$ before obtaining $\frac{d}{dx}H_N^{-1}(x) = 1/H'_N(p_x)$ would be computer-challenging too. Such numerical problems have been tackled by Farouki and his co-authors [18, 16, 17] (see also [3]). Indeed, the central difficulty with Bernstein polynomials lies in the following proposition, proved in [17] :

Proposition 7. *The condition number for the transformation of polynomials of degree N from the Bernstein basis on $t \in [0, 1]$ to the power basis (or vice-versa) in the $\|\cdot\|_1$ and $\|\cdot\|_\infty$ norms is $\kappa(N) \approx 3^{N+1}\sqrt{N+1}/\sqrt{2\pi}$.*

Thus one must avoid to switch between these bases. Consequently, Farouki and Rajan [16] described algorithms for processing polynomials **inside** the Bernstein polynomials system : degree elevation, addition, subtraction, multiplication, division, differentiation, integration, substitution, GCD, resultant of two polynomials, real roots isolation and computation. Even though we only used the direct estimator, such numerical difficulties were met, and resolved by using such algorithms (see Section 4.3.1).

4. Synthesis: the proposed method

To sum up the above facts, the usual Bernstein d.f. and density estimators have contradictory properties. They converge towards well-suited functions, but too slowly : if N is the sample size, we saw in Section 2 that the rate of point-wise convergence of B_N is typically $O\left[\frac{1}{\sqrt{N}}\right]$. But unfortunately, if N is large, because of the numerical problems mentioned above, it is impossible to manage with the associated approximation. Thus, in such cases, it is natural to supersede the step function F_N , supported by the original mesh X_N , by a step function F_K obtained by restricting ourselves to a well-suited sub-mesh X_K of smaller size K . Consider then the “pseudo-Bernstein” approximation corresponding to X_K :

$$\widehat{B}_K[F](x) := \sum_{j=0}^K F(x_j) W_{K,j}(x). \quad (4)$$

It is tempting to supersede (1) by (4), but does it make sense if X_K is not uniform? We will prove that the answer is positive if F is differentiable and if X_K is close enough to uniformity. Suppose now such a sub-mesh has been determined ; in order to improve the fit, we could substitute some Sevy’s operator to B_K , but what number of iterations I^* shall we use in (2)?

Thus, we will face to two problems for processing some d.d.f. F_N :

- find a well-suited sub-mesh of order K such that, simultaneously, K is “reasonable” (e.g. $K \leq 30$), while F_K is a good approximation of F_N (see Sections 4.1 and 4.2)
- choose I^* as large as possible under the constraint that the derivative of $\mathfrak{J}_m^*[F_m]$ is *bona fide* [19] i.e. belongs to both the closed convex cone of positive functions \mathcal{F}^+ and the closed convex set of functions integrating to one, \mathcal{F}^1 (see Section 4.3).

4.1. Optimizing the mesh

We first build a finite sequence of sub-meshes from the initial ordered mesh $X_N := \{0, x_1, \dots, x_{N-1}, 1\}$ and the sequence of uniform meshes $\{U_K : 1 \leq K \leq N\}$; $U_K := \{\frac{i}{K}, 0 \leq i \leq K\}$.

The sub-mesh $X_K := \{0, x_{N_1}, \dots, x_{N_{K-1}}, 1\}$ is such that :

$$\forall 0 \leq m \leq K, x_{N_m} := \arg \min_{x \in X_N} \left| x - \frac{m}{K} \right|.$$

Thus there are often duplicate points in such sub-meshes, which we call degenerate.

4.1.1. Good (sub)meshes via the Hausdorff metric

Suppose F has been sampled on the original mesh X_N , which is not uniform : for some i , $x_i \notin U_N$. The crucial point is that in (1), the value of $F(\frac{k}{m})$ is absolutely required. In their paper, Babu et al. [2] proposed $F_N(\frac{k}{m})$ as an estimator of $F(\frac{k}{m})$ but notice that, unless m divides N , $F_N(\frac{k}{m})$ should be interpolated from the d.d.f. We stress now that the set S_N of jumps of F_N consist in the only experimental points, supported by X_N . For instance, if F_N is a d.d.f.,

$$S_N := \{(x_0 = 0, F(x_0)), (x_1, F(x_1)), \dots, (x_{N-1}, F(x_{N-1})), (x_N = 1, F(x_N))\}.$$

But, for $x \notin X_N$, $F(x)$ could be approached by **any arbitrary** positive increasing function bounded by 1 and interpolating S_N , not especially F_N . That is why we will avoid interpolating the d.d.f., as far as it is possible.

Let us denote $\rho(x)$ the nearest integer function, and introduce the notion of meshes homologous to a uniform one.

Definition 8. We will say that X_K is homologous to U_K if $\forall 1 \leq k \leq K - 1$, $\rho(K x_k) = k$, i.e. the application $\rho(K \bullet) : X_K \mapsto U_K$ is injective.

Suppose first the original mesh X_N is homologous to U_N ; we can write : $\forall k, F(x_k) = F\left(\frac{\rho(N x_k)}{N} + \epsilon_k\right)$, with $\epsilon_k \in]-\frac{1}{N}, \frac{1}{N}[$, and thus $F(x_k) = F\left(\frac{\rho(N x_k)}{N}\right) +$

$\epsilon_k f(\xi_k)$ for some $\xi_k \in \left[\frac{\rho(N x_k)}{N}, \frac{\rho(N x_k)}{N} + \epsilon_k \right]$. Consequently, we have :

$$\begin{aligned} \left| \widehat{B_N}[F](x) - B_N[F](x) \right| &= \left| \sum_{k=0}^N \epsilon_k f(\xi_k) W_{N,k}(x) \right| < \max_{i=0, \dots, N} (|\epsilon_i|) \left| \sum_{k=0}^N f(\xi_k) W_{N,k}(x) \right| \\ &< \frac{1}{N} \left| \sum_{k=0}^N f(\xi_k) W_{N,k}(x) \right| \approx \frac{1}{N} |f(x)|. \end{aligned} \quad (5)$$

Since this desirable situation is generally unfulfilled (if X_N is not uniform, for instance), or N is too large for practical use (see Section 3.2), it seems a good idea to look for a subset of $K-1$ jump positions, $X_K := \{0, x_{N_1}, \dots, x_{N_{K-1}}, 1\}$, such that, for each $k \leq K$, $x_{N_k} \approx \frac{k}{K}$; in this case, the relation (5) will be true, with K instead of N . For this purpose, we need first a measure of distance between X_K and U_K . We consider the Hausdorff distance [11, 35] :

$$d_{\mathcal{H}}(X_K, U_K) = \max \left\{ \max_{0 \leq k \leq K} d(x_k, U_K), \max_{0 \leq k \leq K} d(X_K, k/K) \right\}$$

$$\text{where } d(x_k, U_K) := \min_{0 \leq j \leq K} |x_k - \frac{j}{K}| \text{ and } d(X_K, k/K) := \min_{0 \leq j \leq K} |x_j - \frac{k}{K}|.$$

There is an equivalent definition, based on Minkovski ε -sausages [35] :

Definition 9. Let (E, d) be a metric space, and $P \subset E$. If $B(x, \varepsilon)$ denotes the closed ball of ray ε centered on x , the ε -sausage of P is :

$$P(\varepsilon) := \bigcup_{x \in P} B(x, \varepsilon).$$

The alternative definition of $d_{\mathcal{H}}$ is :

$$d_{\mathcal{H}}(X_K, U_K) = \inf \{ \varepsilon : (X_K \subseteq U_K(\varepsilon)) \wedge (U_K \subseteq X_K(\varepsilon)) \}.$$

Let us now define the resolution of a (sub)mesh.

Definition 10. The resolution of X_K is $R(X_K) := \min_{0 \leq i \leq K-1} \{|x_i - x_{i+1}|\}$.

So, the resolution of degenerate sub-meshes is null. We can now prove the following intermediary issue.

Lemma 11. If $\delta := d_{\mathcal{H}}(X_K, U_K) < \max\left(\frac{1}{K}, \frac{R(X_K)}{2}\right)$, X_K is homologous to U_K .

Proof. Suppose that, under this hypothesis, there exist a pair $(x_k, x_{k'})$ of distinct points such that $\rho(K x_k) = \rho(K x_{k'}) = m$. Then, $x_k = \frac{m}{K} + \epsilon_k$ and $x_{k'} = \frac{m}{K} + \epsilon_{k'}$, with $\max\{|\epsilon_k|, |\epsilon_{k'}|\} < \frac{1}{K}$. Thus, we should have simultaneously $\{x_k, x_{k'}\} \subset B\left(\frac{m}{K}, \delta\right)$ and $|x_k - x_{k'}| \geq R(X_K) > 2\delta$, which is impossible. Thus $\rho(K \bullet) : X_K \mapsto U_K$ must be injective \square

We have finally the following result, much better than (5) :

Theorem 12. *If F is differentiable, we have, under the conditions of Lemma 11 :*

$$\forall x \in [0, 1], \left| \widehat{B}_K[F](x) - B_K[F](x) \right| < d_{\mathcal{H}}(X_K, U_K) |f(x)|.$$

Proof. Immediate from the demonstration of (5), since $\forall 1 \leq k \leq K-1$, $d(x_k, U_K) = d(X_K, \frac{k}{K}) = |\epsilon_k|$ \square

Consequently, when X_K is homologous to U_K , it is potentially a good mesh.

4.1.2. Good (sub)meshes via divergence between binomial distributions

Let's come back to the seminal paper of Bernstein [5], and remember that, since the mode of $\mathcal{B}(N, \frac{i}{N})$ is i , $F(\frac{i}{N})$ has the maximum weight $W_{N,i}(\frac{i}{N})$ in the approximation of $F(\frac{i}{N})$ by (1). Reciprocally, since the mode of $W_{N,k}$ is $\frac{k}{N}$, formula (1) can be written :

$$B_N[F] = \sum_{k=0}^N F(\arg \max (W_{N,k})) W_{N,k}.$$

This stresses how this approximation is deeply linked with Bernoulli trials and the corresponding urn model, studied and generalized by Goldman [20]. Consequently, it is quite natural to measure the closeness of X_K and U_K through distances canonically associated with binomial distributions.

In the light of Informative Geometry [1, 21], the canonical distance between two distributions belonging to some parametric family is given by the geodesic distance between their parameters (Rao distance). In the particular case of $\mathcal{B}(N, a)$ and $\mathcal{B}(N, b)$, it is

$$d_{\mathcal{R}}(a, b) := 2 \arccos \left(\sqrt{ab} + \sqrt{(1-a)(1-b)} \right)$$

(also named Bhattacharyya distance). We will also use the Kullback-Leibler divergence

$$\mathcal{K}(a, b) := a \ln \left(\frac{a}{b} \right) + (1-a) \ln \left(\frac{1-a}{1-b} \right)$$

which is closely linked to Rao distance by geometrical considerations [1, 21]. The Rao distance $d_{\mathcal{R}}(a, b)$ is well-suited for exploratory analyzes while $\mathcal{K}(a, b)$ is closely linked with the likelihood ratio test of the hypothesis : $a \stackrel{?}{=} b$ (prescribed) - see [15]. Indeed, $\mathcal{K}(a, b)$ is asymmetric : it is a divergence, not a distance!

Our two first measures of global divergence of X_K from U_K (wished) are then

$$\overline{d_{\mathcal{R}}}(X_K, U_K) := \max_{1 \leq m < K} d_{\mathcal{R}} \left(x_{N_m}, \frac{m}{K} \right)$$

from the one hand, and

$$\overline{\mathcal{K}}(X_K, U_K) := \max_{1 \leq m < K} \mathcal{K} \left(x_{N_m}, \frac{m}{K} \right).$$

from the other hand.

We introduce too a geometrical criterion, based on $d_{\mathcal{R}}$. Consider the vector

$$\overrightarrow{\Psi(X_K)} := \frac{1}{K} \left(\sqrt{x_{N_1}}, \sqrt{1-x_{N_1}}, \dots, \sqrt{x_{N_{K-1}}}, \sqrt{1-x_{N_{K-1}}} \right).$$

Since it belongs to the unit sphere of $\mathbb{R}^{2(K-1)}$, as well as $\overrightarrow{\Psi(U_K)}$, a natural distance between X_K and U_K is the arc length

$$\mathcal{A}(X_K, U_K) := \arccos \left(\overrightarrow{\Psi(X_K)} \cdot \overrightarrow{\Psi(U_K)} \right) < \frac{\pi}{2}.$$

We can now propose a synthetic definition for good meshes, alias G-meshes, as local Pareto optima of the considered criteria.

Definition 13. A sub-mesh X_{K^G} is a G-mesh if the four criteria $d_{\mathcal{H}}(X_K, U_K)$, $\overline{d_{\mathcal{R}}}(X_K, U_K)$, $\overline{\mathcal{K}}(X_K, U_K)$ and $\mathcal{A}(X_K, U_K)$ simultaneously present a local minimum for $K = K^G$.

4.2. Optimal sub-meshes for d.d.f.s

Generally, several G-meshes stem from the original mesh X_N , but notice that they are independent from the studied d.d.f. or, equivalently, its set of jumps S_N . For instance, the reader can see on Figure 1 that we had to select a single mesh from five G-meshes. That is why we propose an additional criterion for selecting an optimal sub-mesh, taking into account the whole data set, and not only the mesh : if F_{K^G} denotes the d.d.f. corresponding to the sub-set of jumps located on X_{K^G} , we will choose $K^G < N$ as small as possible, in order that the Hausdorff distance $d_{\mathcal{H}}(F_{K^G}, F_N)$ is small too (see Figure 2).

The choice of this distance is supported by the works of Beer [4] and Cuevas and Fraiman [11] : $d_{\mathcal{H}}$ is a metric in the space of Upper Semi Continuous functions, and any d.f. is USC. Furthermore, if the theoretical d.f. F is continuous, the propositions $\|F_K - F\| \xrightarrow{K \rightarrow \infty} 0$ and $d_{\mathcal{H}}(F_K, F) \xrightarrow{K \rightarrow \infty} 0$ are equivalent [4].

Remark. Since F_{K^G} and F_N are step functions, $d_{\mathcal{H}}(F_{K^G}, F_N)$ only depends on the position of the vertices of the completed graphs [11] V_N and V_{K^G} of these functions. For instance, if F_N is a e.d.f.

$$V_N := S_N \cup \left\{ \left(x_{(1)}, 0 \right), \left(x_{(2)}, \frac{1}{N} \right), \dots, \left(x_{(N-1)}, \frac{N-2}{N} \right), \left(x_{(N)} = 1, \frac{N-1}{N} \right) \right\}.$$

One can easily prove that $\forall K \leq N$, $d_{\mathcal{H}}(F_K, F_N) = d_{\mathcal{H}}(V_K, V_N)$, and the last distance is much easier to compute than the first one.

4.2.1. A benchmark in Density Estimation : the explosions in mines data

These data consist in the time-interval (in days) between severe explosions in British mines, between 6 December 1875 and 29 May 1951. They were extensively studied by Simonoff [33], after Boneva et al. [6]. We start with a uniform discretization of these data into 55 classes ([33], p.222).

All the criteria proposed in Section 4.1 are represented on Figure 1. On each panel of this figure, seven points are put in evidence with markers : the only G-meshes. Six of these points, labeled by a black diamond, are associated with sub-meshes homologous to the corresponding uniform mesh ; they correspond to $K^G \in \{3, 6, 7, 13, 27, 55\}$. The seventh one ($K^G = 23$) is labeled by an empty circle. Choosing 3 or 55 is unreasonable, but what about 13 or 27 knots? Indeed, one can clearly see on Figure 2 that $K^* = 27$ seems the best choice.

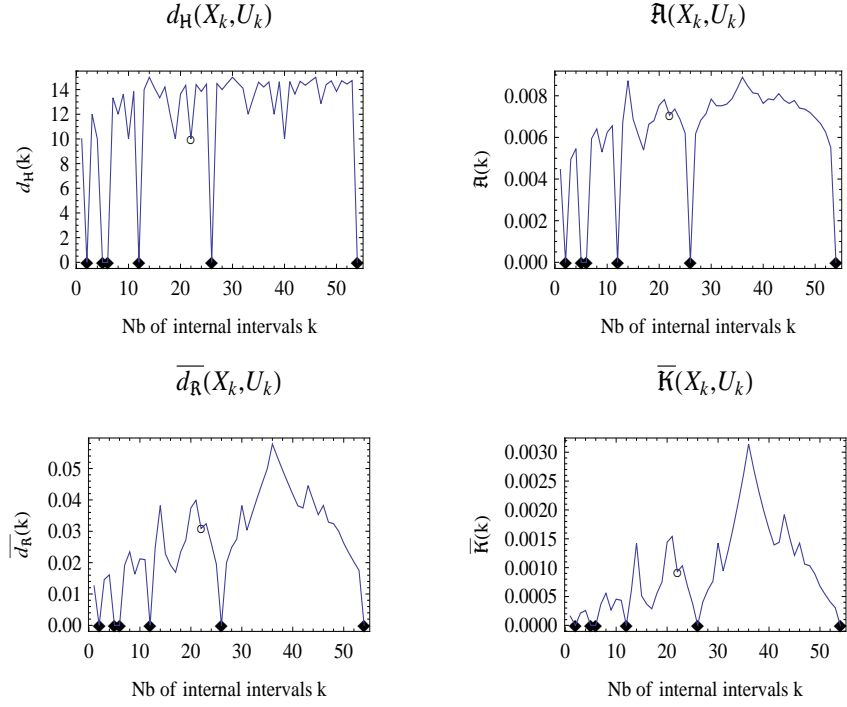


Figure 1:

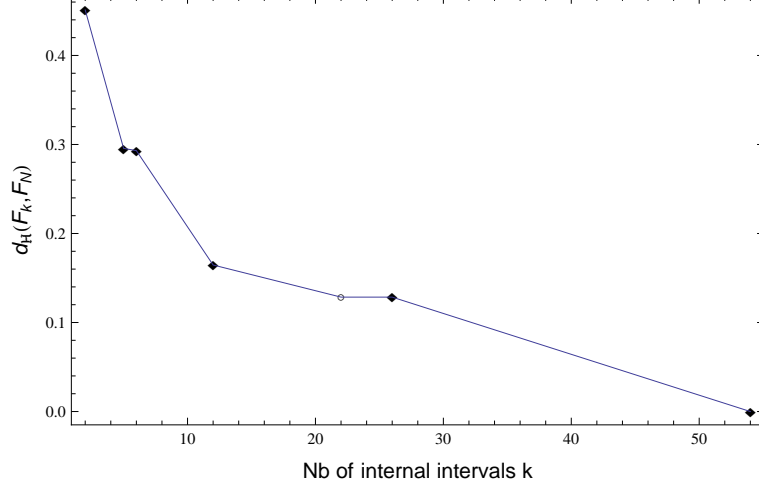


Figure 2:

4.3. Optimizing the number of iterations

We will suppose from now that F is differentiable with derivative f , and that we could determine an optimal G-mesh X_{K^*} homologous to U_{K^*} ; for simplicity of notations, we will fix here : $K = K^*$.

In accordance with Theorem 12, the approximation (4) is close to (1), and $d_H(F_K, F_N)$ is small. Since K is small too, we will improve the approximation of F by $\widehat{B}_K[F]$ through Sevy's iterations. We show hereunder that a *bona fide* approximation [19] of f can be obtained this way.

Let us first introduce an *ad hoc* d.d.f. \tilde{F} , such that $\forall 0 \leq k \leq K$, $\tilde{F}(\frac{k}{K}) = F(x_k)$; then $\widehat{B}_K[F] = B_K[\tilde{F}]$. While $B_K[\tilde{F}]$ is also a d.f., due to the shape-preserving properties of Bernstein operator, it is likely that $\lim_{I \rightarrow \infty} \mathfrak{J}_K^I[\tilde{F}] = L_K[\tilde{F}]$ has no special structure, since Lagrange interpolation is unconstrained. Furthermore, the error $\|L_K[\tilde{F}] - \tilde{F}\|$ can grow exponentially with K in our case, because “uniform spacing of data can have bad consequences” (de Boor [13, Ch. 2]; see also Laurent [22, Ch. 5]), due to the unboundedness of Lebesgue's functions associated with such meshes.

Thus, both of these approximations have bad properties. Searching for a trade-off, Cooper and Waldron [10] proposed to run across the whole segment $tB_K[\tilde{F}] + (1-t)L_K[\tilde{F}]$, $0 \leq t \leq 1$. We will follow a slightly different line, specific to density approximation. Notice that the trajectory $\left\{ \widehat{f}_K^{(i)} = \frac{d\mathfrak{J}_K^i[\tilde{F}](x)}{dx}, 1 \leq i \right\}$ consists in a sequence of functions starting in $\mathcal{F}^+ \cap \mathcal{F}^1$ ($\widehat{f}_K := \frac{d\widehat{B}_K[F](x)}{dx} = \frac{dB_K[\tilde{F}](x)}{dx} = \frac{d\mathfrak{J}_K^1[\tilde{F}](x)}{dx}$ is *bona fide*) and progressively getting out this closed convex set (in general, $\frac{d\mathfrak{J}_K^\infty[\tilde{F}](x)}{dx} = \frac{dL_K[\tilde{F}](x)}{dx} \notin \mathcal{F}^+ \cap \mathcal{F}^1$). Consequently, it is

quite natural to select the first I^* such that $\widehat{f_K}^{(I^*)}$ belongs to $\mathcal{F}^+ \cap \mathcal{F}^1$, while $\widehat{f_K}^{(I^*+1)}$ doesn't. For that purpose, we will control the graph of $\widehat{f_K}^{(i)}$ through two "stresses" : the positivity one

$$\pi(i) := \int_0^1 \left(\left| \widehat{f_K}^{(i)} \right| - \widehat{f_K}^{(i)} \right) (x) dx \quad (6)$$

and the normality one

$$\nu(i) := \int_0^1 \left(\widehat{f_K}^{(i)} + \left| \widehat{f_K}^{(i)} \right| \right) (x) dx - 2. \quad (7)$$

The estimator $\widehat{f_K}^{(i)}$ is *bona fide* if and only if both these stresses are null.

Remark. Since in our case $\int_0^1 \left| \widehat{f_K}^{(i)} \right| (x) dx \approx 1$, the curves π and ν are very similar to each other.

4.3.1. Computational aspects

Let us denote \mathfrak{B}_K the linear space generated by $\{W_{K,j} : 0 \leq j \leq K\}$, $\partial_K \in \mathcal{L}(\mathfrak{B}_K, \mathfrak{B}_{K-1})$ the differentiation operator, and $\sigma_K \in \mathcal{L}(\mathfrak{B}_K, \mathfrak{B}_{K+1})$ the integration operator. Using well-known results [20, 16], one can easily verify that the matrix of ∂_K relatively to Bernstein bases is the $(K, K+1)$ band matrix :

$$\begin{pmatrix} -K & K & 0 & \cdots & 0 \\ 0 & -K & K & 0 & \vdots \\ \vdots & 0 & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & -K & K \end{pmatrix}$$

and that $\text{Ker}(\partial_K)$ is the unit vector of \mathbb{R}^{K+1} (because of the partition of unity property : $\sum_{k=0}^K W_{n,k}(x) = 1$). We can thus write :

$$\widehat{f_K} = \frac{dB_K[\tilde{F}](x)}{dx} = \partial_K \circ \tilde{F}_K = K \sum_{i=0}^{K-1} W_{K-1,i} \left(\tilde{F} \left(\frac{i+1}{K} \right) - \tilde{F} \left(\frac{i}{K} \right) \right)$$

which is positive because \tilde{F} is increasing and all the $W_{K-1,i}$ are positive.

Reciprocally, the matrix of σ_K relatively to Bernstein bases is the $(K+2, K+1)$ lower-triangular matrix :

$$\begin{pmatrix} 0 & \cdots & 0 \\ \frac{1}{K+1} & \ddots & \vdots \\ \vdots & \ddots & 0 \\ \frac{1}{K+1} & \cdots & \frac{1}{K+1} \end{pmatrix}.$$

One can easily show that $\partial_{K+1} \circ \sigma_K = Id_{K+1}$, while

$$\sigma_K \circ \partial_{K+1} = Id_{K+2} - \begin{pmatrix} 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \cdots & 0 \end{pmatrix}.$$

Consequently, $\sigma_{K-1} \circ \widehat{f_K} = \widetilde{F}_K - \widetilde{F}(0) \cdot \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$, and $\int_0^1 \widehat{f_K}(x) dx = (\sigma_{K-1} \circ \widehat{f_K})(K) = \widetilde{F}(1) - \widetilde{F}(0) = 1$ (this show once more that $\widehat{f_K}$ is *bona fide*).

Notice now that any polynomial $\mathfrak{B}_{K-1} \ni P := \sum_{i=0}^K p_i W_{K-1,i}$ can be decomposed in a non-negative component $P^+ := \sum_{i=0}^K \{p_i W_{K-1,i} : p_i \geq 0\}$ and a negative component $P^- := \sum_{i=0}^K \{p_i W_{K-1,i} : p_i < 0\}$.

From $\widehat{f_K}^{(i)} = \widehat{f_K^+}^{(i)} + \widehat{f_K^-}^{(i)}$, we get :

$$\int_0^1 \widehat{f_K}^{(i)}(x) dx \leq \int_0^1 \left| \widehat{f_K}^{(i)} \right|(x) dx \leq \int_0^1 \widehat{f_K^+}^{(i)}(x) dx - \int_0^1 \widehat{f_K^-}^{(i)}(x) dx,$$

which can be re-written in an alternative form :

$$\left(\sigma_{K-1} \circ \widehat{f_K}^{(i)} \right)(K) \leq \int_0^1 \left| \widehat{f_K^+}^{(i)} \right|(x) dx \leq \left(\sigma_{K-1} \circ \widehat{f_K^+}^{(i)} \right)(K) - \left(\sigma_{K-1} \circ \widehat{f_K^-}^{(i)} \right)(K). \quad (8)$$

Formulas (6) and (7) can be re-written too in a simpler form :

$$\pi(i) = \int_0^1 \left| \widehat{f_K}^{(i)} \right|(x) dx - \left(\sigma_{K-1} \circ \widehat{f_K}^{(i)} \right)(K) \quad (9)$$

and

$$\nu(i) = \left(\sigma_{K-1} \circ \widehat{f_K}^{(i)} \right)(K) + \int_0^1 \left| \widehat{f_K}^{(i)} \right|(x) dx - 2. \quad (10)$$

While computing the first right member of these formulas is easy and numerically stable [16], computing $\int_0^1 \left| \widehat{f_K}^{(i)} \right|(x) dx$ can be a problem when one has

to transform $\widehat{f}_K^{(i)}$ to the power form for integration, because of Proposition 7. In such cases, the inequation (8) is very useful to point numerical problems.

4.3.2. Back to the data

The frequency of explosions in mines

We plotted on Figure 3 both the stresses (9) and (10), together with the Kolmogorov distance (in percents) $K.D.(i) := 100 \sup_{x \in X_N} \left| \int_0^x \widehat{f}_K^{(i)}(t) dt - F_N(x) \right|$, for $i \leq 7$. The reader can see on this figure that this function steeply decreases until $i = 3$, and then remains approximately constant around 0.018%. According to this figure, I^* must be chosen in $[3, 7]$, since both stresses are very small ($< 10^{-3}\%$) ; we chose 6, which corresponds to the minimum Kolmogorov distance (0.018 %).

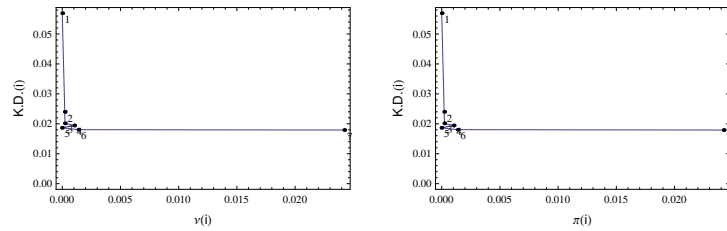


Figure 3:

The obtained approximations are plotted on Figure 4. The reader can compare our results to those of Simonoff [33], who tested on these data six density estimation methods (kernel (p. 49-50), boundary kernel (p. 53-54), local likelihood (p.68, 242-243), logsplines (p.71), local polynomial regression (p. 223-224) and Loess (p. 240)), or those of Boneva et al. [6], who used the histospline.

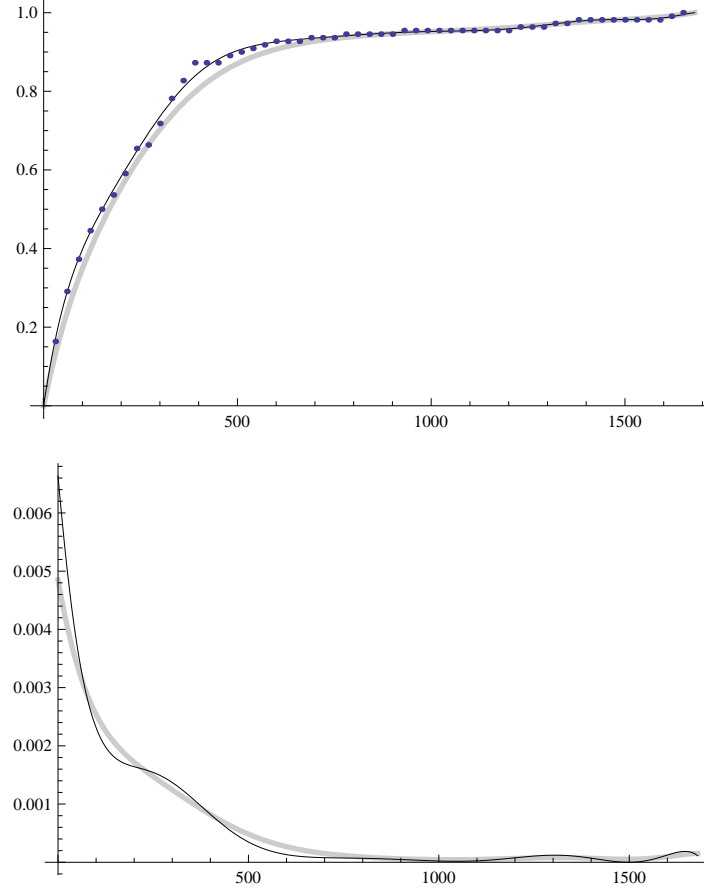


Figure 4:

This exemplifies the excellent behavior of our approximation near the boundary, probably linked with the results of Bouezmarni and Rolin [8] ; indeed, if the occurrence of such explosions approximately obeys a Poisson process, it is natural that the densities plotted on Figure 4 are close to an exponential distribution, as noticed by Boneva et al. [6].

The Bulgarian maternity data

These data can be found in Boneva et al. [6] (p.23), and consist in the number of maternities in Bulgaria in 1963, per age class of the mother. There were no birth to women younger than 15 years or older than 50 years. The complete curve has $N = 36$ points spread on a uniform mesh, but our method led us to select a sufficient sub-mesh of $K^* = 16$ points. One can see on Figure 5 that in this case the stress curves are not L-shaped as in Figure 3 : both stresses grow until $i = 4$ and then decrease until $i = 10 = I^*$.

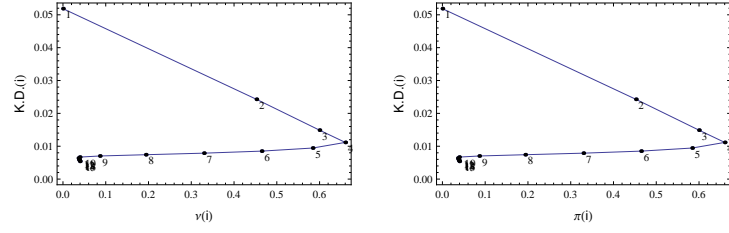


Figure 5:

The approximated distribution functions and densities are shown on Figure 6, which can be compared with Figure 9 of [6] (p. 69).

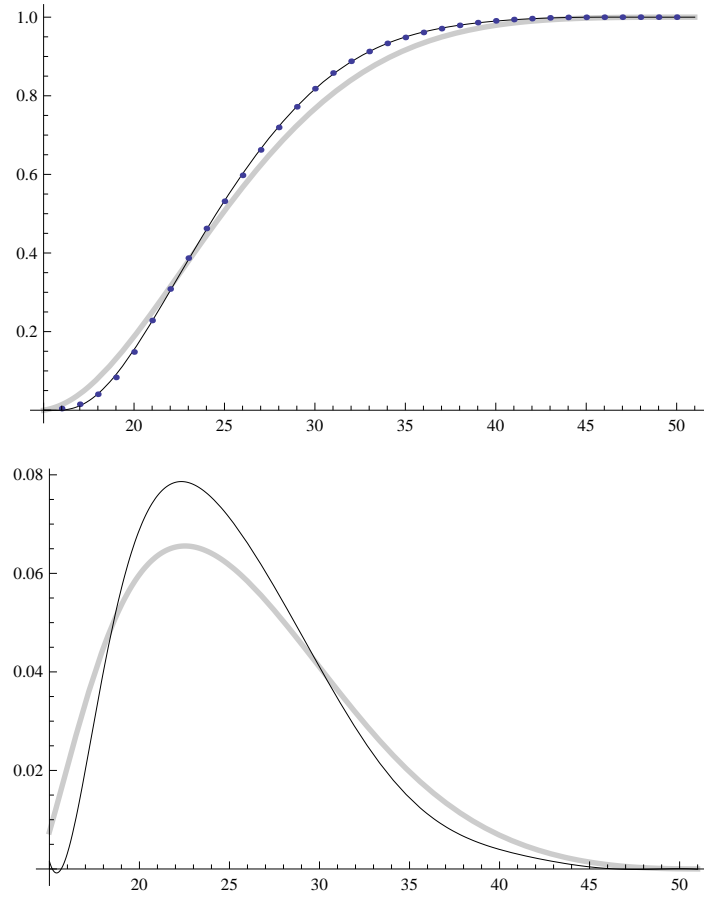


Figure 6:

Our approximation is far better than the classical Bernstein's one, and doesn't show any "rabbit", contrary to the histospline of [6]. But it shares several interesting features of the histospline : its positive skewness, and small negative ordinates for mothers around 15 years. According to L. Boneva [6, p. 69], this is not a flaw : this is due to truncation of the data, since mothers younger than 15 years were not investigated.

Sedimentology : processing a reference grain-size curve

In order to investigate the impact of macrobenthic fauna activity on sedimentary structures, an experiment was carried out, giving rise to a set of 552 grain-size curves. These curves were discretized according to a common mesh of size $N = 92$, which was not uniform. Our goal here is to find, at last approximately, the correct number and positions of fractiles of one of these curves (a control sediment), denoted F_μ . The associated d.d.f. is plotted on Figure 7. Following the same strategy as in previous examples, we first selected an optimal sub-mesh ($K^* = 30$), and then a convenient number of iterations ($I^* = 3$).

Consider now the obtained approximation of F_μ , $\mathfrak{P} := \mathfrak{J}_{K^*}^{I^*}[\tilde{F}]$. It is a polynomial of degree 29, strictly increasing since $\widehat{f_{K^*}}^{(I^*)} = \frac{d\mathfrak{P}}{dx}$ is positive. Thus, for any $M \leq N$, we can compute the vector of fractiles :

$$\mathfrak{Q}(M) := \left(\mathfrak{P}^{-1}\left(\frac{1}{M}\right), \dots, \mathfrak{P}^{-1}\left(\frac{M-1}{M}\right) \right)$$

and compare it to its closest sub-mesh of size M extracted from X_N , denoted Q_M (notice that $Q_M \neq X_M$ is often degenerate, while $\mathfrak{Q}(M)$ is not). Notice that here, Q_M plays the part that U_M played in Section 4.1.

Plotting the Hausdorff distances $\{d_{\mathcal{H}}(X_3, Q_3), \dots, d_{\mathcal{H}}(X_N, Q_N)\}$ (after eliminating duplicate abscissa) and the Hausdorff distances between d.d.f.s (after eliminating duplicate points), as in section 4.2, we found that $M^* = 17$ is likely the best choice. The crude data and both approximations of F_μ are represented on Figure 7, while both densities are plotted on Figure 8.

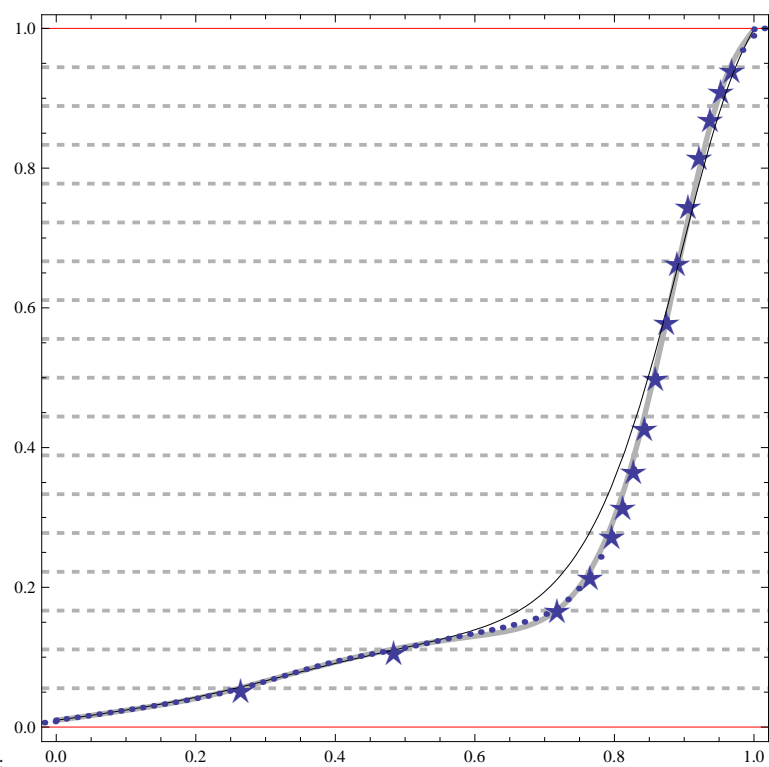


Figure 7:

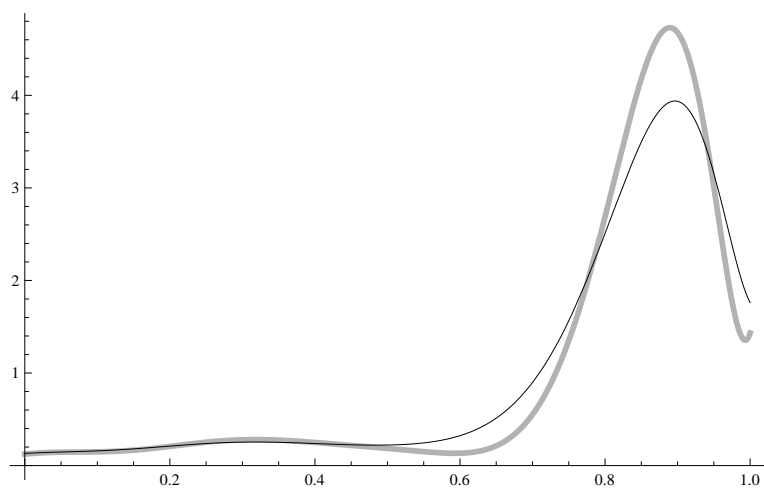


Figure 8:

5. Discussion

We emphasized in the introduction the dual belonging of Bernstein approximation to the fields of Approximation Theory and Statistics ; it will be also invoked here. We propose in this work a two-step method for density approximation. The first step consists in putting ourselves in position to use the Bernstein operator, by choosing an optimal G-mesh X_{K^*} . In the second one, we select a number of iterations I^* such that the density approximation $\widehat{f}_K^{(I^*)}$ belongs to $\mathcal{F}^+ \cap \mathcal{F}^1$, while $\widehat{f}_K^{(I^*+1)}$ is outside.

From a statistical perspective, the optimization of I^* could be slightly different if F_N is an e.d.f. Remember that, for $I \leq I^*$, each $\mathcal{J}_K^I[\tilde{F}] \in \mathfrak{B}_K$ has the structure of a d.f. parametrized by $K+1$ coordinates. Consequently, we can consider in this case the Kolmogorov distance $\sup_{x \in X_N} |\mathcal{J}_K^I[\tilde{F}](x) - F_N(x)|$ as a supplementary stopping test for I . This cannot be done straightforwardly, because the $K+1$ coordinates of $\mathcal{J}_K^I[\tilde{F}]$ must be estimated, but this function as well as the d.f. can be separately estimated on two different random sub-samples, and the goodness-of-fit tested [34]. Another testing method would be to estimate $\mathcal{J}_K^I[\tilde{F}]$ on the first half-sample, and compute $U_I := \left\{ \mathcal{J}_K^I[\tilde{F}](x_j), 1 \leq j \leq [N/2] \right\}$ on the second half-sample. If the fit is good, U_I obeys the standard uniform law, and this hypothesis can be easily checked by Hilbertian tests (see for instance [7, 23]).

From an Approximation perspective, notice that the heart of the method is quite versatile : the main point is that $F \in \mathfrak{C}$, where \mathfrak{C} is a convex set of continuous functions, preserved by the Bernstein operator. Here, we were interested in the set $\mathcal{F}^+ \cap \mathcal{F}^1$ and the stresses used were associated with that \mathfrak{C} . But we could proceed the same way with other constraints. Suppose for instance \mathfrak{C} is the set of convex functions supported by $[0, 1]$ and differentiable twice : in this case, the relevant stress would be $c(i) := \int_0^1 \left(\widehat{F}_K'' - \left| \widehat{F}_K'' \right| \right)(x) dx$ since $\widehat{F}_K'' := \partial_{K-1} \circ \partial_K \circ \mathcal{J}_K^I[\tilde{F}]$ should be positive.

Acknowledgement

The author is very grateful to Drs. F. Gilbert (ECOLAB laboratory, UMR CNRS 5245, Toulouse, France) and G. Stora (LMGEM) for providing the grain-size data.

Figures captions

Figure 1: Plot of the four quality criteria for sub-meshes. See comments in the text.

Figure 2: The Hausdorff distance between d.d.f.s in the case of G-meshes.

Figure 3: Plot of both stresses versus Kolmogorov distance for the Mines data, as parametric curves of the iterations number I .

Figure 4: Plot of the approximate d.f. and density for Mines data. In gray : Bernstein approximation ; in black : approximation with the iterated operator ($I^* = 6$). Abscissas are expressed in days.

Figure 5: Plot of both stresses versus the Kolmogorov distance for the maternity data, as parametric curves of the iterations number I .

Figure 6: Plot of the approximate d.f. and density of births relative to mother's age. In gray : Bernstein approximation ; in black : approximation with the iterated operator ($I^* = 10$). Abscissas are expressed in years.

Figure 7: Plot of the approximate d.f.s associated with the reference grain-size curve (dotted curve). Thin black curve : the Bernstein approximation ; in gray : approximation with the iterated operator ($I^* = 3$). The stars correspond to the d.d.f. F_{M^*} associated with Q_{M^*} , while the horizontal grid gives the position of the fractiles of both approximations.

Figure 8: Plot of the approximate grain-size densities. Thin black curve : the Bernstein approximation ; in gray : approximation with the iterated operator.

References

- [1] S. Amari, Differential-geometrical methods in Statistics. Lecture Notes in Statistics 28, Springer-Verlag, Berlin, 1985.
- [2] G. J. Babu, A. J. Canty and Y. P. Chaubey, Application of Bernstein polynomials for smooth estimation of a distribution and density function. *Journal of Statistical Planning and Inference*, 105(2002), 377-392.
- [3] R. Barrio and J. M. Peña, Evaluation of the derivative of a polynomial in Bernstein form. *Applied Mathematics and Computation*, 167(2005), 125-142.
- [4] G. Beer, Upper semicontinuous functions and the Stone approximation theorem. *Journal of Approximation Theory*, 34(1982), 1-11.
- [5] S. N. Bernstein, Démonstration du théorème de Weierstrass fondée sur le calcul des probabilités. *Commun. Soc. Math. Kharkov*, 13(1912), 1-2.
- [6] L. I. Boneva, D. Kendall and I. Stefanov, Spline transformations : three new diagnostic aids for the statistical data-analyst (with discussion). *Journal of the Royal Statistical Society, Ser. B*, 33(1971), 1-71.
- [7] D. Bosq, Tests du χ^2 généralisés. Comparaison avec le test du χ^2 classique. *Revue de Statistique Appliquée*, XXXVII, 1(1989), 43-52.
- [8] T. Bouezmarni, and J.M. Rolin, Bernstein estimator for unbounded density function. *Journal of Nonparametric Statistics*, 19, 3(2007), 145-161.
- [9] C. Chui, Concerning rates of convergence of Riemann sums. *Journal of Approximation Theory*, 4(1971), 279-287.
- [10] S. Cooper and S. Waldron, The eigenstructure of the Bernstein operator. *Journal of Approximation Theory*, 105(2000), 133-165.
- [11] A. Cuevas, R. Fraiman, On visual distances in density estimation : the Hausdorff choice. *Statistics & Probability Letters*, 40(1998), 333-341.
- [12] P. J. Davis, Interpolation and approximation. Blaisdell, New York, 1963.

- [13] C. de Boor, A practical guide to splines. Applied Mathematical Sciences, 27, Springer-Verlag, New York, 1978.
- [14] R. de Bruin, D. Salomé and W. Schaafsma, A semi-Bayesian method for nonparametric density estimation. Computational Statistics & Data Analysis, 30(1999), 19-30.
- [15] S. Egushi and J. Copas, Interpreting Kullback-Leibler divergence with the Neyman-Pearson lemma. Journal of Multivariate Analysis, 97(2006), 2034-2040.
- [16] R. T. Farouki, V. T. Rajan, Algorithms for polynomials in Bernstein form. Computer Aided Geometric Design, 5(1988), 1-26.
- [17] R. T. Farouki, On the stability of transformations between power and Bernstein polynomials forms. Computer Aided Geometric Design, 8(1991), 29-36.
- [18] R. T. Farouki, Convergent inversion approximations for polynomials in Bernstein form. Computer Aided Geometric Design, 17(2000), 179-196.
- [19] L. Gajek, On improving density estimators which are not bona fide functions. The Annals of Statistics, 14, 4(1986), 1612-1618.
- [20] R. N. Goldman, Urn models, approximations, and splines. Journal of Approximation Theory, 54(1988), 1-66.
- [21] R. E. Kass, The geometry of asymptotic inference. Statistical Science, 4, 3(1989), 188-234.
- [22] P.-J. Laurent, Approximation et optimisation. Enseignement des sciences, 13, Hermann, Paris, 1972.
- [23] T. Ledwina, Data-driven version of Neyman's smooth test of fit. Journal of the American Statistical Society, 89, 427(1994), 1000-1005.
- [24] C. Manté, The use of regularization methods in computing Radon-Nikodym derivatives. Application to grain-size distributions. SIAM Journal on Scientific Computing, 21, 2(1999), 455-472.

- [25] C. Manté, J.P. Durbec, C. Degiovanni and F. Werth, Mapping grain-size data through multivariate analysis and regression. Comparison of kriging, loess and splines, in: S.J. Lippard, A. Naess and R. Sinding-Larsen (Eds), Proceedings of IAMG'99, Trondheim, Norway, 6-11 August 1999, Vol. 1, pp. 75-80.
- [26] C. Manté, J.P. Durbec and J.C. Dauvin, A functional data-analytic approach to the classification of species according to their spatial dispersion. Application to a marine macrobenthic community from the Bay of Morlaix (Western english Channel). *Journal of Applied Statistics*, 32, 8(2005), 831-840.
- [27] C. Manté, A.F. Yao and C. Degiovanni, Principal component analysis of measures, with special emphasis of grain-size curves. *Computational Statistics & Data Analysis*, 51(2007), 4969-4983.
- [28] S. Petrone, Random Bernstein Polynomials. *Scandinavian Journal of Statistics*, 26(1999), 373-393.
- [29] A. Sahai, An iterative algorithm for improved approximation by Bernstein's operator using statistical perspective. *Applied Mathematics and Computation*, 149(2004), 327-335.
- [30] J. C. Sevy, Convergence of iterated boolean sums of simultaneous approximants. *Calcolo*, 30(1993), 41-68.
- [31] J. C. Sevy, Lagrange and least-squares polynomials as limits of linear combinations of iterates of Bernstein and Durrmeyer polynomials. *Journal of Approximation Theory*, 80(1995), 267-271.
- [32] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*, Chapman & Hall, London, 1992.
- [33] J. S. Simonoff, *Smoothing methods in Statistics*, Springer Series in Statistics, New York, 1998.
- [34] M. A. Stephens, On the half-sample method for goodness-of-fit. *Journal of the Royal Statistical Society, Series B*, 40, 1(1978), 64-70.
- [35] C. Tricot, *Courbes et dimensions fractales*, Springer-Verlag, Paris, 1993.

- [36] X. Zeng, A. Piriou, On the rate of convergence of two Bernstein-Bézier type operators for bounded variation functions. *Journal of Approximation Theory*, 95(1998), 369-387.